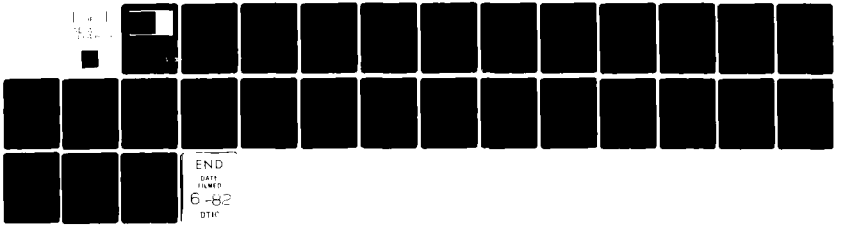
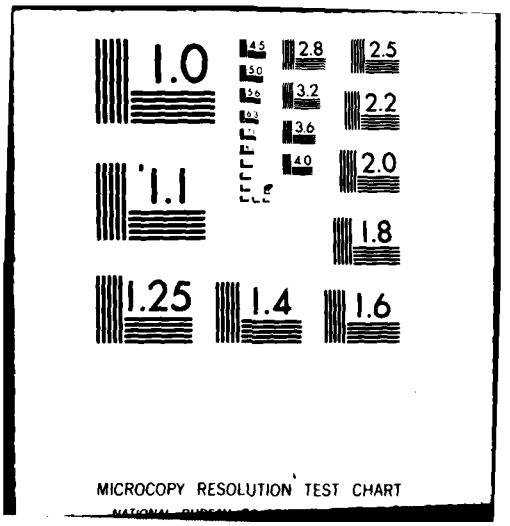
AD-A114 623 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
SOME PENALIZED LIKELIHOOD PROCEDURES FOR SMOOTHING PROBABILITY --ETC(U)
FEB 82 T ATILGAN, T LEONARD DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2336 NL



END
DATE FILMED
6-82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART

MRC Technical Summary Report #2336

SOME PENALIZED LIKELIHOOD PROCEDURES FOR SMOOTHING PROBABILITY DENSITIES

Taskin Atilgan and Tom Leonard

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

February 1982

(Received January 7, 1982)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

SOME PENALIZED LIKELIHOOD PROCEDURES FOR SMOOTHING PROBABILITY DENSITIES

Taskin Atilgan and Tom Leonard

Technical Summary Report #2336

February 1982

ABSTRACT

Some methods are considered for the estimation of probability densities. They employ a linear approximation to either the density or the logistic density transform. The coefficients in the approximation are estimated by maximum likelihood and the number of terms is judged via an information criterion. Hence the traditionally difficult problem of judging the degree of smoothness is carried out in a relatively simple manner. Criteria considered include the penalties proposed by Akaike and Schwarz for model complexity, together with an empirical criterion based upon a plot of the log-likelihoods. The practical procedures are related to an asymptotic consistency argument and a number of numerical examples are presented.

AMS (MOS) Subject Classification: 62G05

Key Words: Non-parametric, Density estimation, Histogram, Information Criteria, Consistency.

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

A

SIGNIFICANCE AND EXPLANATION

Non-parametric methods for the estimation of probability densities traditionally experience severe technical and conceptual difficulties in judging the appropriate degree of smoothness for the estimated density. In this paper some relatively simple procedures are discussed which employ either polynomial or histogram approximations and refer to an information criterion to judge the appropriate number of parameters in the approximation. The duality between parameter parsimony and smoothness is therefore exploited. Information criteria discussed include those due to Akaike and Schwarz together with an empirical criterion based upon a plot of the log-likelihoods. The practical procedures are motivated by an asymptotic consistency argument and a number of numerical examples are presented.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

SOME PENALIZED LIKELIHOOD PROCEDURES FOR SMOOTHING PROBABILITY DENSITIES

Taskin Atilgan and Tom Leonard

1.1 Parameter Parsimony and Density Estimation

Consider the estimation of an unknown density $f(x)$, for $x \in (a,b)$ from a random sample of observations $x_1,\ldots,x_n$. We will employ approximations for $f(x)$ taking the form

$$f_m(x) = f_m(x|\underset{\sim}{\theta}_m) = \lambda\{\underset{\sim}{\theta}_m \underset{\sim}{\psi}_m(x)\} = \lambda\{\sum_{j=1}^{m} \theta_j \phi_j(x)\} \qquad (1.1.1)$$

where $\lambda(\cdot) : R \to R$ is a known transformation, $\underset{\sim}{\theta}_m = (\theta_1,\ldots,\theta_m)$ is a vector of unknown coefficients, and $\underset{\sim}{\psi}_m = (\phi_1,\ldots,\phi_m)^T$ corresponds to a basis of specified functions, for example

(a) $\underset{\sim}{\psi}_m(x) = (1,x,x^2,\ldots,x^{m-1})$

(b) $\phi_1,\ldots,\phi_m$, a suitable set of orthonormal polynomials

(c) $\phi_j(x) = I_{T_j}(x)$ for $j = 1,\ldots,m$, where the I_{T_j} are indicator functions for a partition $(T_1,\ldots,T_m)$ of the sample space. When $\lambda(\cdot)$ is the identity function this leads to the histogram situation discussed in section 2.1.

(d) $\phi_1,\ldots,\phi_m$ corresponding to an appropriate set of B-splines, so that $\underset{\sim}{\theta}_m \underset{\sim}{\psi}_m(x)$ gives a general formulation of a cardinal spline with $m - 3$ knots (our ideas on splines will be developed more fully in another paper)

(e) $\phi_1,\ldots,\phi_m$ representing terms in a Fourier Series

The vector $\underset{\sim}{\theta}_m$ will be typically estimated by maximization of its log-likelihood

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

$$L_m(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \log \lambda\{\underset{\sim m}{\theta}\, \psi_{\sim m}(x_i)\} \quad . \tag{1.1.2}$$

Let $\hat{\underset{\sim}{\theta}}$ denote the maximum likelihood vector and $L_m = L_m(\hat{\underset{\sim}{\theta}})$ the maximum value achieved by the likelihood. These quantities will also be important in judging an appropriate value for m; one possibility is to apply information criteria proposed by Akaike (1974), and Schwarz (1978), for model selection, to this situation.

Akaike and Schwarz suggest choosing the model which maximizes the respective penalized log-likelihoods

$$AIC = L_m - m \tag{1.1.3}$$

and

$$SIC = L_m - \frac{1}{2} m \log n \quad . \tag{1.1.4}$$

These criteria introduce penalties for large values of m and therefore enforce "parameter parsimony." In the present context, parameter parsimony can be viewed as representing smoothness of the unknown density f. The penalties could therefore be referred to as "roughness penalties" in the manner of I. J. Good (see section 1.3).

Akaike obtained AIC by an intuitive argument based upon entropy measures. The alternative, SIC, has been frequently suggested in the Bayesian literature e.g. by Jeffreys (1961), in a variety of special cases. A fairly general approximate Bayesian formulation has been suggested by Leonard (1981) and Chow (1981). The modification

$$SIC' = L_m - \frac{1}{2} m \log(n/2\pi) \tag{1.1.5}$$

might be better for moderate sample sizes.

The three criteria so far described are special cases of the general criterion

$$GIC = L_m - \alpha m \tag{1.1.6}$$

where α is the penalty per parameter in the model. In this paper we will

consider AIC, SIC, SIC' together with an empirical procedure EIC described below for choosing α.

Note that under fairly general regularity conditions the log-likelihood L_m will possess the following asymptotic behaviour as $n \to \infty$

$$\begin{aligned} \frac{1}{n} L_m &\overset{a.s.}{\to} \sup_{\underset{\sim}{\theta}_m} \int f(x) \log f_m(x|\underset{\sim}{\theta}_m)dx \\ &= \int f(x) \log \tilde{f}_m(x)dx \\ &= \eta_m(f,\tilde{f}_m) \quad , \end{aligned} \tag{1.1.7}$$

where

$$\eta_m(f,\tilde{f}_m) = -I(f,\tilde{f}_m) + \int f(x) \log f(x)dx \tag{1.1.8}$$

with

$$\tilde{f}_m(x) = \sup f_m(x|\tilde{\underset{\sim}{\theta}}_m) \tag{1.1.9}$$

where $\tilde{\underset{\sim}{\theta}}_m$ achieves the supremum in (1.1.7) and

$$I(f,\tilde{f}_m) = -\int f(x) \log\{\tilde{f}_m(x)/f(x)\}dx \tag{1.1.10}$$

denotes the Kullback-Liebler information distance between f and $\tilde{f}_m$. Note that as $n \to \infty$ maximization of $n^{-1}L_m$ with respect to m corresponds to minimization of this information distance.

Consider, for illustrative purposes, the polynomial situation described above where $\psi_m(x) = (1,x,x^2,\ldots,x^{m-1})$ and suppose that the true density takes the form

$$f(x) = f_{m^*}(x) = \lambda\{\underset{\sim}{\theta}_{m^*}\underset{\sim}{\psi}_{m^*}(x)\} \tag{1.1.11}$$

where m^* is the "true" number of terms in the linear approximation. Since the functional form of f is often unspecified in practice, it will often make sense to simply refer to this special form involving m^*.

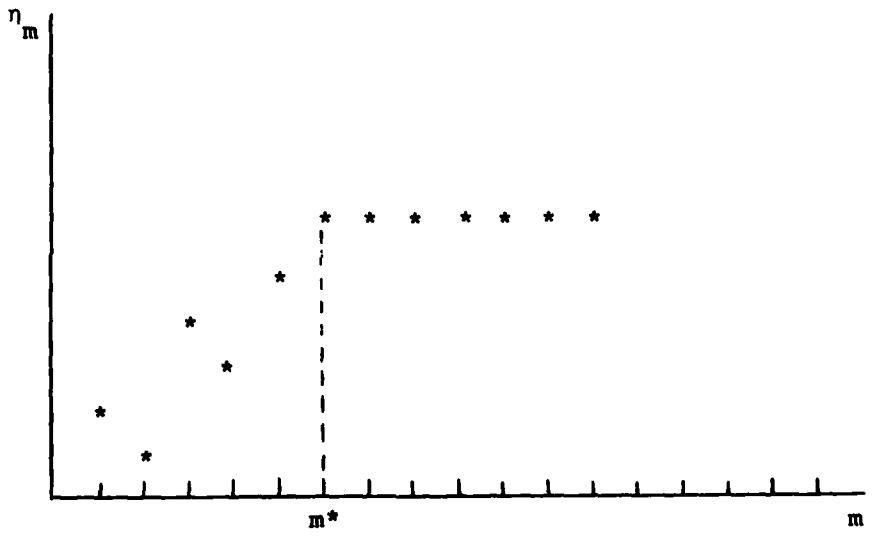
In this case, $\eta(f,\tilde{f}_{m^*})$ achieves the maximum over m of the quantities in (1.1.8), with zero Kullback information distance, yielding perfectly consistent estimation, and we also have

$$\eta_m(f, \tilde{f}_m) = \eta_{m^*}(f, \tilde{f}_{m^*}) \qquad (1.1.12)$$

$$\text{for} \quad m = m^*, m^* + 1, \ldots.$$

since the extra coefficients for $\underset{\sim}{\theta}_m$ with $m > m^*$ are set equal to zero in (1.1.9). Therefore the η_m will follow the type of behaviour depicted in Figure 1.

Figure 1: Limiting behaviour of log-likelihoods as $n \to \infty$

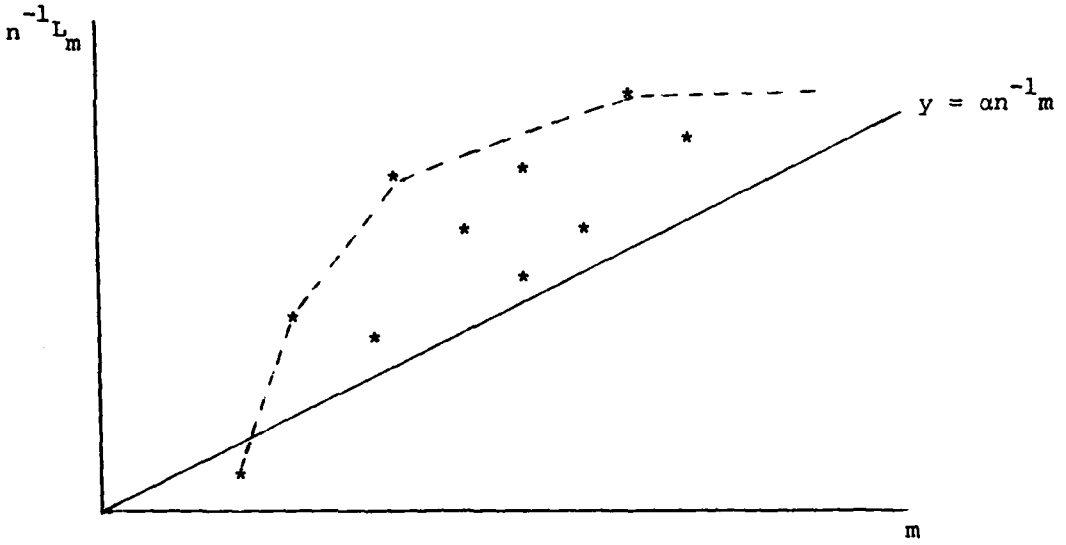


Note that η_m will tend to rise with m until it achieves the ridge at $m = m^*$, when the points flatten out. Therefore the problem at issue is how to infer this ridge point from the observed quantities $n^{-1}L_m$. We omit detailed asymptotic results which show that as $n \to \infty$, both AIC and SIC give values of m which almost surely satisfy $m \geq m^*$, and lead to consistent estimators for the true density in (1.1.10). Since SIC is based upon an

asymptotically larger penalty than AIC this suggests that SIC may lead to values of m which are on the whole closer to m^* than those suggested by AIC.

The observed $n^{-1}L_m$ will, for finite n, behave like the theoretical points in Figure 1, but with random fluctuation around these points (see Figure 2).

Figure 2: Typical Behaviour of Log-Likelihoods for Finite n



(---- is the upper convex boundary)

When the penalty for parameter parsimony is α m, maximization of GIC leads to the value of m such that the $(m, n^{-1}L_m)$ point in Figure 2 maximizes the perpendicular distance above the line $y = \alpha n^{-1} m$. Therefore the upper convex boundary in Figure 2 contains at its vertices all points which are optimal according to GIC, for some value of α. For example, the optimal AIC and SIC points must both correspond to a vertex of this boundary.

The upper convex boundary restricts attention to a convenient subset of values of m. Sometimes it may be possible to judge the ridge-point m^* in Figure 1 by simple visual inspection. Alternatively empirical rules can be devised which estimate the position of the ridge point e.g. by empirically choosing a value of α rather than referring to SIC or AIC. An example (EIC) of such a rule involves minimization with respect to α and m_α of

$$\varepsilon = \alpha m_\alpha \tag{1.1.13}$$

where $m_\alpha > 1$ is a value of m which maximizes

$$GIC = L_m - \alpha m \quad . \tag{1.1.14}$$

Minimization of the product in (1.1.13) creates a compromise between fidelity to the data, as represented by a small value of α, and smoothness of the estimate for f, represented by a small value for m. We are intuitively speaking trying to obtain maximum smoothing for minimum penalty αm. Whilst this procedure is ad hoc, our asymptotic argument described in the next paragraph suggests that it gives a value of m converging almost surely to the true value m^*. Also, for finite n, it seems to give reasonable answers in empirical situations.

The EIC criterion gives an optimal value which may be obtained graphically (see Figure 2). We should use the vertex of the convex boundary which minimizes m times the slope of the line segment of the boundary lying immediately to the right of m. For the theoretical values in Figure 1 we see that this automatically leads to the true value m^*, where we assume $m^* > 1$. Since the $n^{-1}L_m$ converge almost surely to the η_m in Figure 1, this suggests, unrigorously, that EIC gives a consistent estimator for m^*.

The ideas discussed above may also be applied to the histogram and spline situations (c) and (d) introduced at the beginning of this section, but with some modifications to the limiting behaviour of the log-likelihoods. These will be discussed, for histogram smoothing, in section 2.1.

One way of measuring the performances of AIC, SIC, SIC', and EIC is via simulated observations from a "true" density f. We will judge the difference between the estimated and true values by quantities associated with the Kullback-Liebler information distance in (1.1.10).

Two contributions to the literature slightly related to our approach are by Crain (1974), and Geisser and Eddy (1979).

1.2 Further Uses of Parameter Parsimony

The procedures described in section 1.1 depend upon the choice of basis vector $\underset{\sim}{\psi}_m(x)$. However they may also be employed for basis section e.g. for deciding whether to use, say, a Fourier Series approximation rather than polynomials. It is simply necessary to compare the suprema of the penalized log-likelihoods for the different basis selections. An example will be presented in section 2.3.

All of our criteria may be used for estimating any one-dimensional function via a linear approximation e.g. hazard functions for survival data, the logit function in non-parametric bioassay, and possible non-linear regression functions.

For the standard linear statistical model $y_j = \underset{\sim}{\theta} \underset{\sim}{x}_j + \varepsilon_j$ for $j = 1,\ldots,m$, where the ε_j are uncorrelated normal errors with constant variance σ^2, GIC reduces to the minimization of

$$\frac{n}{2} \log(n\, R_m) + \alpha m$$

where R_m is the residual sum of squares. An entropy based criterion

proposed by Box and Kanamasu (1973) uses a similar adjustment, but with $\alpha = \frac{1}{2}$ to account for model complexity. It is also possible to show that $\alpha = 1$ (AIC) is equivalent to Mallows' C_p (see Mallows, 1973).

1.3 Other related work

More formal approaches for the smooth estimation of curves have been proposed by Wahba (1977), Good and Gaskins (1971), and Leonard (1978), leading to more sophisticated roughness penalties. For example, Wahba effectively modifies the log-likelihood by

$$-\frac{1}{2}\lambda \int (f^{(2)}(t))^2 dt \quad .$$

The parameter λ controls the degree of smoothing and plays a similar role to m under the present approach. This parameter is usually estimated via cross-validation, but the computations become particularly difficult in non-linear situations. An information criterion for selecting m seems to be much simpler, and seems to answer a traditinally difficult problem by estimating the degree of smoothness in a relatively simple manner.

Table 1: Values of Criteria for Example 1

m	L_m	$L_m - \frac{m-1}{2}$	AIC	SIC	SIC'
1	.00	.00	-1.00	-2.30	-1.38
2	2.01	1.51	.01	-2.59	-.75
3	1.27	.27	-1.73	-5.63	-2.87
4	2.76	1.26	-1.24	-6.44	-2.76
5	13.44	11.44	8.44	1.94	6.54
6	4.76	2.26	-1.24	-9.04	-3.52
7	10.23	7.23	3.23	-5.87	.57
8	9.47	5.97	1.47	-8.93	-1.57
9	11.85	7.85	2.85	-8.85	-.57
10	14.84	10.34	4.84	-8.16	1.84
11	12.68	7.68	1.68	-12.62	-2.50
12	11.65	6.15	-.35	-15.95	-4.91
13	15.17	9.17	2.17	-14.73	-2.77
14	11.79	5.29	-2.21	-20.41	-7.53
15	16.30	9.30	1.30	-18.20	-4.40
16	17.12	9.62	1.12	-19.68	-4.96
17	13.77	5.77	-3.23	-25.33	-9.69
18	18.46	9.96	.46	-22.94	-6.38
19	14.30	5.30	-4.70	-29.40	-11.92
20	18.91	9.41	-1.09	-27.09	-8.69
21	17.29	7.29	-3.71	-31.01	-11.69

2. APPLICATIONS

2.1 Choosing the Interval Width for a Grouped Histogram

Let $x_1, x_2, \ldots, x_n$ represent a random sample of raw observations from a distribution with density f concentrated, for convenience, on $(0,1)$. We will estimate f by a grouped histogram with m equal intervals, and will judge the smoothness of the histogram by estimating m and hence the interval width m^{-1}. This contrasts with a procedure proposed by Leonard (1973) for smoothing the probabilities in a histogram with fixed intervals. Let $n_1, \ldots, n_m$ denote the cell counts when the n observations are grouped into m intervals.

Our histogram estimator has the form

$$\hat{f}_m(x) = m \sum_{j=1}^{m} P_j \, I_{T_j}(x) \qquad (2.1.1.)$$

where P_j is the proportion of raw observations lying in interval $T_j = (m^{-1}(j-1), m^{-1}j]$, for $j = 1, \ldots, m$, and $I_{T_j}(x)$ is the appropriate indicator function. The log-likelihood for fixed m is then

$$\begin{aligned} L_m &= \sum_{i=1}^{n} \log \hat{f}_m(x_i) \\ &= n \log m + \sum_{j=1}^{n} n_j \log P_j \, . \end{aligned} \qquad (2.1.2)$$

Utilizing a result due to Basharin (1959) we have, as $n \to \infty$,

$$E(n^{-1}L_m) = \eta_m(f, \tilde{f}_m) + \frac{(m-1)}{2n} + O\left(\frac{m}{n^2}\right) \qquad (2.1.3)$$

where

$$E(\eta_n(f, \tilde{f}_m)) = \int f(x) \log \tilde{f}_m(x) dx \qquad (2.1.4)$$

and $\tilde{f}_m$ is defined in a similar manner to the quantity in (1.1.9).

Therefore $(m-1)/2n$ can be viewed as the "bias" when $n^{-1}L_m$ is used to estimate the expectation in (2.1.5). Note that closeness of $n^{-1}L_m - (m-1)/2n$ to the true value in (2.1.5) is an indication that the Kullback-Liebler information distance in (1.1.10) is close to zero i.e. it indicates that this value of m gives an estimate of f which is close to f according to this distance measure.

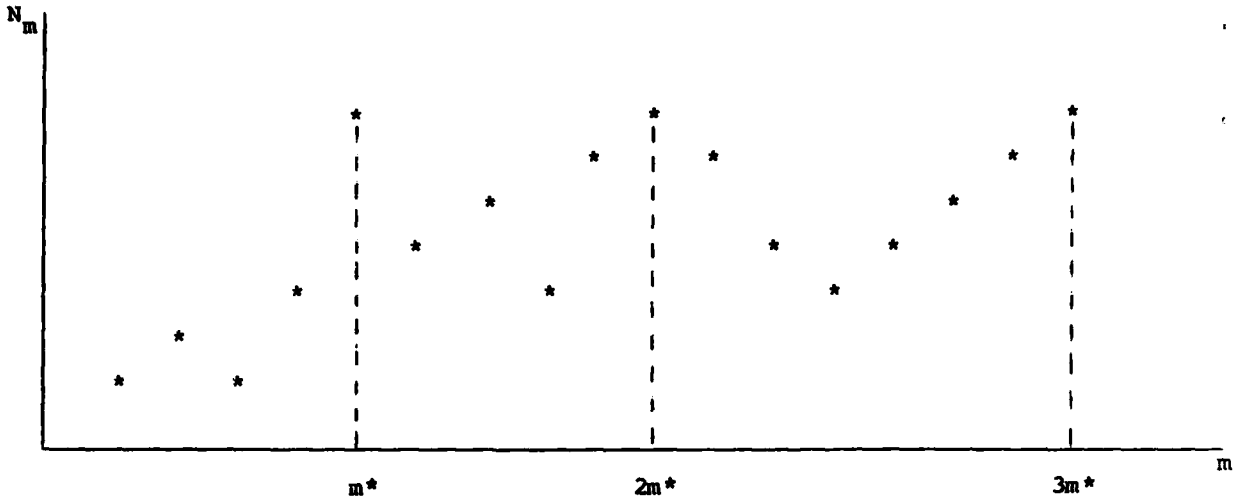
Example 1: We generated $n = 100$ observations from a "true" histogram with $m = m^* = 5$ and all probabilities 0.125, 0.25, 0.125, 0.375, and 0.125.

With this choice of f the Kullback-Liebler information distance in (1.1.10) is of course minimized when $m = 5$. It yields a theoretical value of 0.1153 for the expression in (2.1.4).

The simulation generated the empirical probabilities 0.11, 0.25, 0.9, 0.38 when $m = 5$. Note from the 3rd column of Table 1 that the bias-corrected log-likelihood assumes the value 11.44, very close to the maximum theoretical value possible of 11.53. This confirms tht we would like our selection procedures to choose $m = 5$ in this situation. Note, from the 3rd to 6th columns, that AIC, SIC, and SIC' are all clearly maximum when $m = 5$. We have checked that this is true for all $m = 1,2,\ldots,100$. The three criteria respectively correspond, under GIC, to $\alpha = 1,2,3$ and 1.38.

Suppose in general that the true density f is an equal interval histogram with $m = m^*$ cells. Then the expression in (2.1.4) will be maximized when $m = m^*$ and also when $m = 2m^*, 3m^*, \ldots$. Therefore, in contrast to the limiting behaviour in Figure 1, for the polynomial situation, the $n^{-1}L_m$, and also the $n^{-1}L_m - (m-1)/2n$, will possess the limiting behaviour depicted in Figure 3.

Figure 3: Limiting behaviour of histogram log-likelihoods as $n \to \infty$



Again the empirical criterion, EIC, defined in (1.1.13) and (1.1.14) will lead to a consistent estimator for m^* as $n \to \infty$.

For our simulation study, the log-likelihoods are plotted in Figure 4 for $m = 1,2,\ldots,20$; they were however calculated for all values of m up to 100. The vertices on the upper boundary are at $m = 5$, 64, and 99. A visual inspection suggests that $m = 5$ lies at the beginning of the ridge. According to GIC, the choice $m^* = 5$ is optimal when α lies between 0.4 and 0.8. Higher choices are disregarded as too extreme to achieve optimality for a reasonable value of α.

The primary choice, between $m = 5$ and $m = 64$, can be made by reference to EIC, comparing $0.8 \times 5 = 4.0$ with $0.4 \times 64 = 25.6$; again suggesting that $m^* = 5$ is best.

Figure 4: Plot of Log-Likelihoods

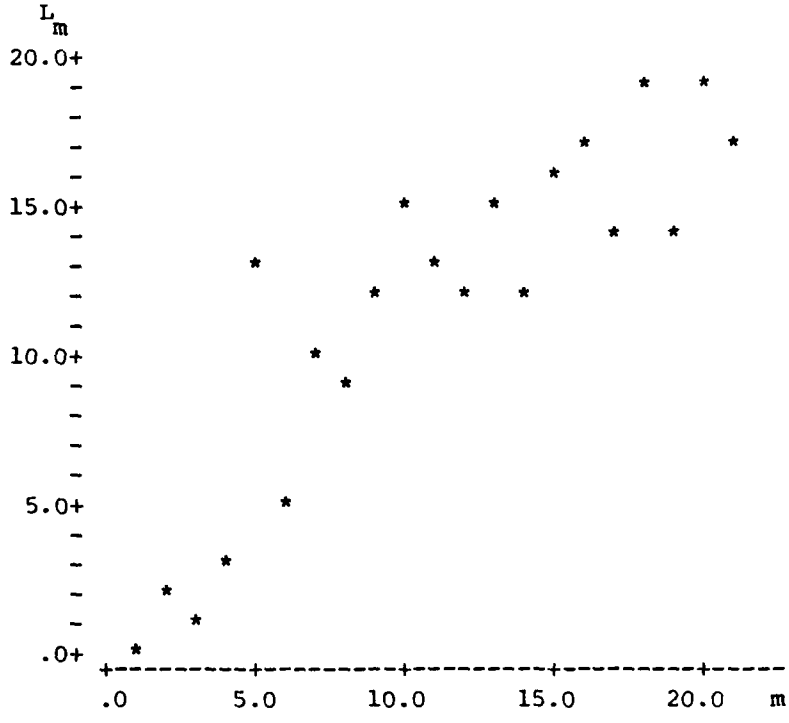
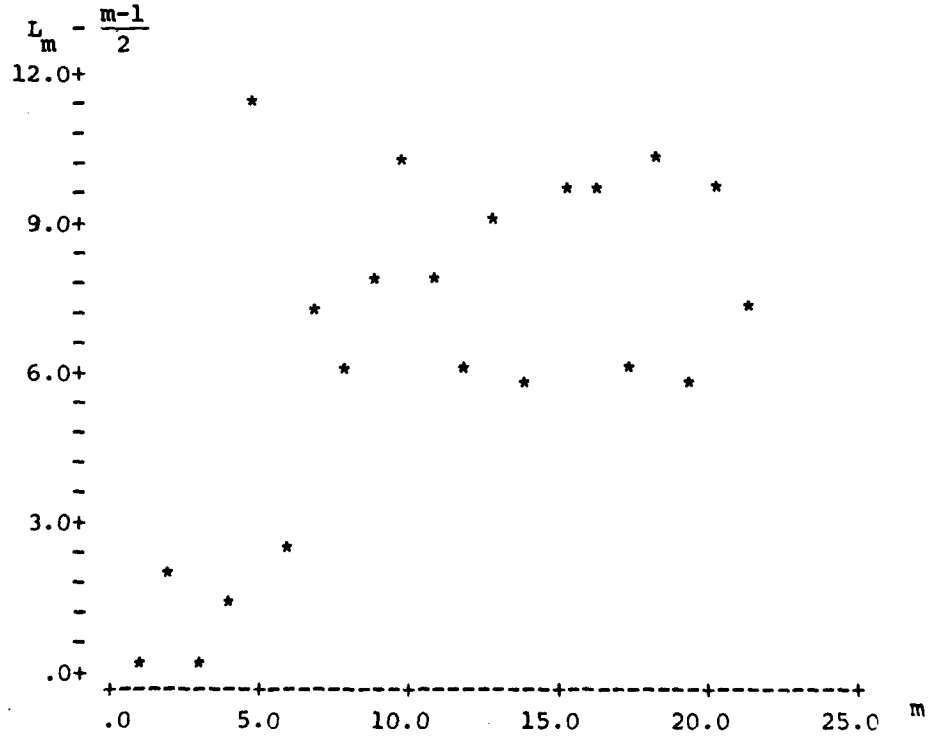


Figure 5: Plot of Bias-adjusted Log-Likelihoods



The statistician might prefer to carry out a similar procedure but based on the plot of the bias-adjusted log-likelihoods in Figure 5. Note that direct maximization amongst these adjusted terms does not in general work since it tends towards higher values of m than produced via an information criterion. The latter again give $m = 5$ as best.

Example 2: We next generated 500 random observations from a mixture $0.3N(-1,1) + 0.7N(6,1)$ of two normal distributions. The results are described in Table 2 and Figure 6. The plot of bias-adjusted log-likelihoods is fairly similar to Figure 6 and is therefore omitted. A visual inspection of Figure 6 suggests that any value of m between 11 and 167 would probably be reasonable. Note that the quantity in (2.1.4) assumes the value 1013/500 when $\hat{f}_m$ is replaced by the true density f. Comparison with the bias corrected log-likelihoods in Table 2 suggests that $m = 17$ is best, but only marginally superior than $m = 11$.

In this case both AIC and EIC gave $m = 17$ and SIC and SIC' gave $m = 11$. In Figure 7 the two corresponding histograms are compared with the true density. It seems that $m = 17$ gives just the right amount of smoothing.

Table 2: Values of Criteria for Example 2

m	L_m	$L_m - \frac{m-1}{2}$	AIC	SIC	SIC'
1	-1319	-1319	-1320	-1322	-1321
2	-1276	-1277	-1278	-1282	-1281
3	-1284	-1285	-1287	-1293	-1291
4	-1250	-1252	-1254	-1263	-1259
5	-1137	-1139	-1142	-1153	-1148
6	-1109	-1112	-1115	-1128	-1122
7	-1090	-1093	-1097	-1112	-1105
8	-1067	-1071	-1075	-1092	-1085
9	-1067	-1071	-1076	-1095	-1087
10	-1048	-1053	-1058	-1079	-1070
11	-1019	-1024	-1030	-1053	-1043
12	-1040	-1046	-1052	-1077	-1066
13	-1026	-1032	-1039	-1067	-1055
14	-1037	-1044	-1051	-1081	-1068
15	-1036	-1043	-1051	-1083	-1069
16	-1020	-1028	-1036	-1070	-1055
17	-1007	-1015	-1024	-1060	-1044
18	-1022	-1031	-1040	-1078	-1062
19	-1022	-1031	-1041	-1081	-1064
20	-1017	-1027	-1037	-1080	-1061
21	-1015	-1025	-1036	-1080	-1061
22	-1007	-1017	-1029	-1075	-1055
23	-1016	-1027	-1039	-1087	-1066
24	-1004	-1016	-1028	-1079	-1057
25	-1009	-1021	-1034	-1087	-1064
26	-1008	-1020	-1034	-1089	-1065
27	-1006	-1019	-1033	-1090	-1065
28	-1008	-1022	-1036	-1095	-1070
29	-1013	-1027	-1042	-1103	-1076
30	-1005	-1019	-1035	-1098	-1071

Figure 6: Plot of Log-Likelihoods

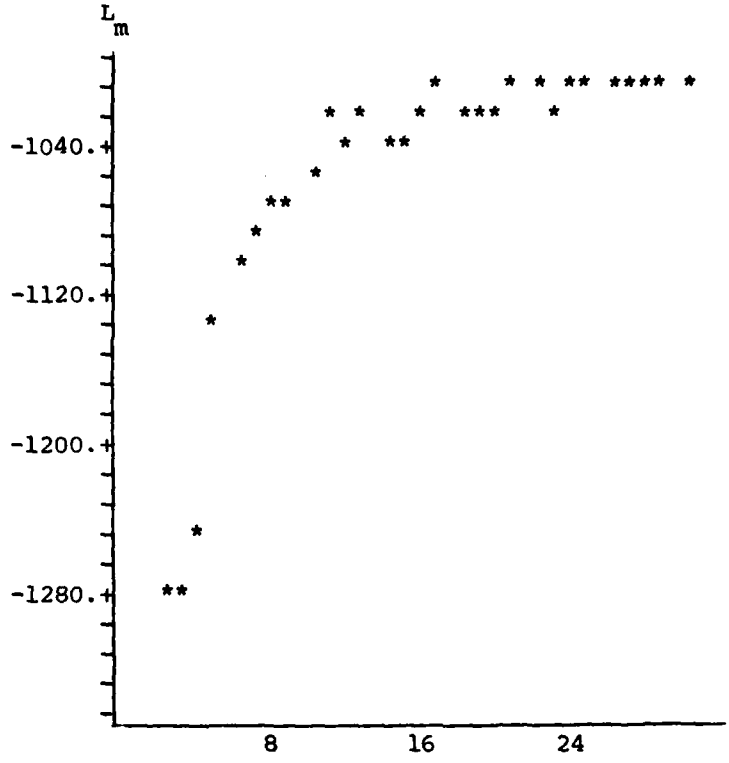
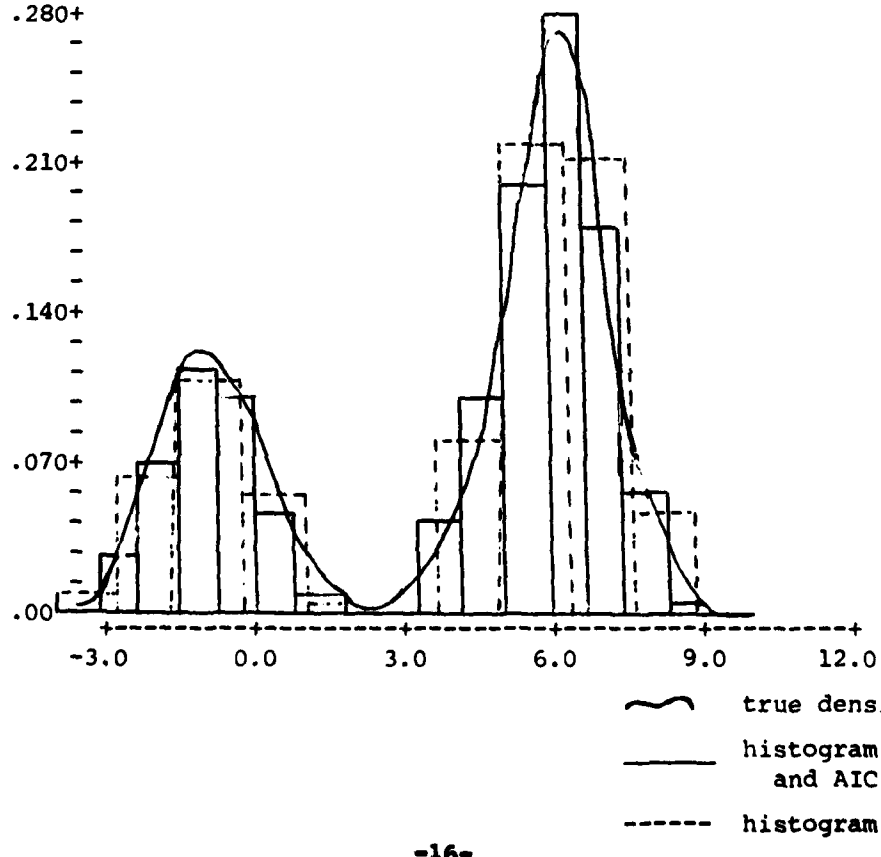


Figure 7: Comparison of Histograms and True Density



Example 3: We next consider the well-known chondrite meteor data collected by Ahrens (1965), and previously analyzed by Leonard (1978), Silverman (1981), and Good and Gaskins (1980).

Table 3: Silica Assays of 22 Chondrite Meteors

.04	.15	.16	.78	.40	.44	.45	.46	.49	.54
.92	.59	.64	.75	.81	.33	.84	.85	.87	.87

The log-likelihood plot in Figure 8 yields the choice $m = 16$ under EIC, with $\alpha = 0.85$, a value also selected by SIC' with $\alpha = 0.63$. However both AIC and SIC choose $m = 5$ with $\alpha = 1$ and $\alpha = 1.55$ respectively. The histograms for $m = 5$ and $m = 16$ are depicted in Figures 9 and 10.

Both histograms suggest trimodality of the underlying distribution, agreeing with the results of Good and Gaskins. However $m = 16$ seems to create a preferable balance between smoothness and fidelity to the data.

For Examples 1 - 3 we see that EIC gives the most reasonable results in all three situations while AIC, SIC, and SIC' are more variable in performance. In the next section EIC again fares well.

Figure 8: Log-Likelihoods for Chondrite Meteor Data

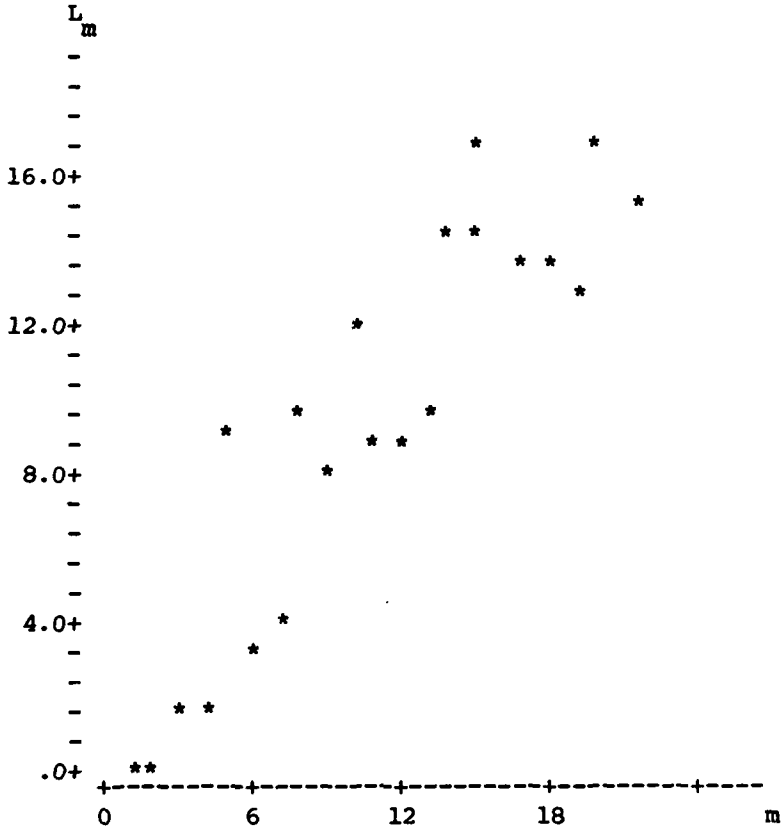


Figure 9: (m = 5, AIC, SIC)

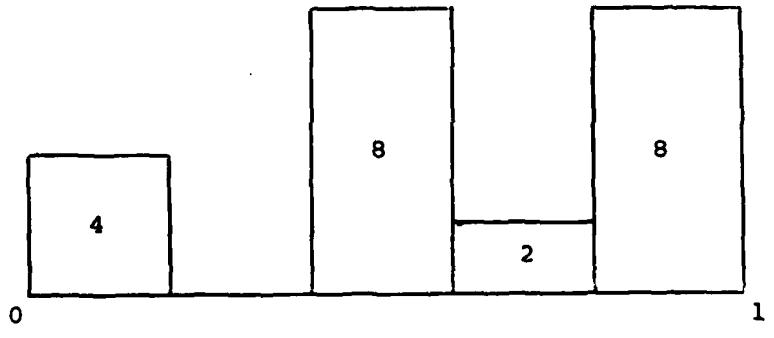
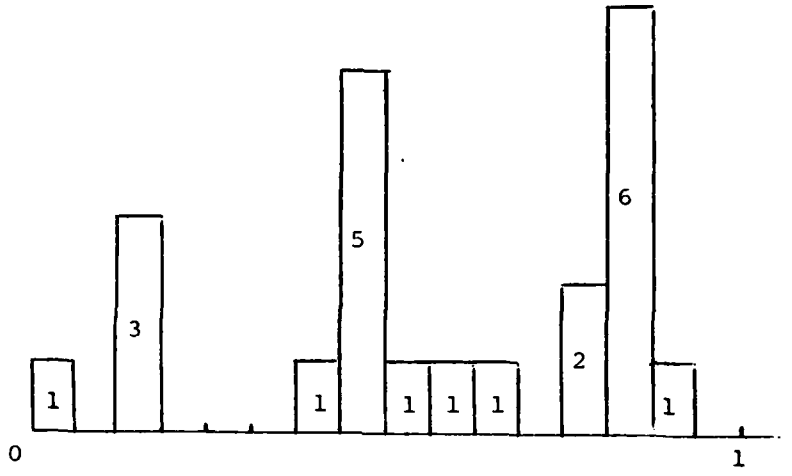


Figure 10: (m = 16, SIC', EIC)



2.2 Logistic Density Estimation

Apart from the histogram situation it is not usually satisfactory to take $\lambda(\cdot)$ in (1.1.1) to be the identity function since this creates non-negativity constraints on $\underset{\sim}{\theta}_m$ and it is difficult to either estimate $\underset{\sim}{\theta}_m$ by maximum likelihood or to ensure that the density integrates to unity. However, it is sometimes reasonable to follow Leonard (1978) by considering the logistic form

$$f_m(\underset{\sim}{x}|\underset{\sim}{\theta}_m) = \frac{e^{\underset{\sim}{\theta}_m \underset{\sim}{\psi}_m(x)}}{\int_b^b e^{\underset{\sim}{\theta}_m \underset{\sim}{\psi}_m(s)} ds} \quad (a < x < b) \qquad (2.2.1)$$

as an approximation to the underlying density f. For example, the choice $\underset{\sim}{\psi}_m(x) = (x, x^2, \ldots, x^m)^T$ ensures that the first m sample moments constitute a set of sufficient statistics for $\underset{\sim}{\theta}_m$. Transformation to a set of suitable orthogonal polynomials yields well-conditioned Hessians for the iterative

solution of the likelihood equations by Newton-Raphson. The latter involves maximization of the log-likelihood function

$$L_m(\underset{\sim}{\theta}_m) = \underset{\sim}{\theta}_m \underset{\sim}{t}_m - n \log \int_b^b \exp\{\underset{\sim}{\theta}_m \underset{\sim}{\psi}_m(x)\}dx \qquad (2.2.3)$$

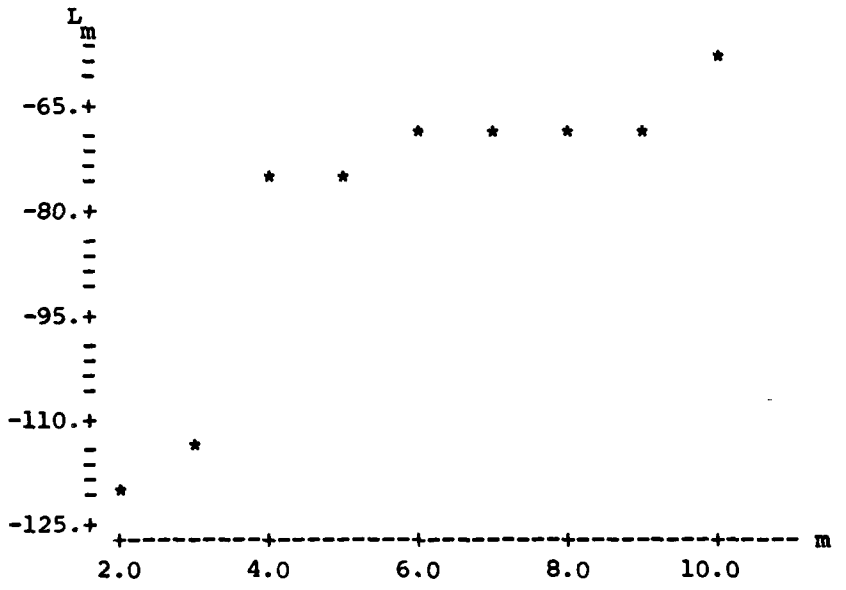
where

$$\underset{\sim}{t}_m = \sum_{i=1}^{n} \underset{\sim}{\psi}_m(x_i) \qquad (2.2.4)$$

and can be tackled using standard optimization packages.

In our numerical example we completed this maximization for $m = 1,2,\ldots,10$ (by which time we had obviously reached the ridge-point) and then selected m via our information criteria.

Figure 11: Log-Likelihoods for Logistic Example



Example: 200 random numbers are generated from the mixture $0.3N(-1,1) + 0.7N(6,1)$ of two normal distributions. Using a basis of polynomials up to degree m we calculated the log-likelihoods depicted in Figure 11.

A visual inspection suggests that either $m = 4$ or $m = 6$ is appropriate. However EIC, SIC, SIC', and AIC respectively chose $m = 4$, 6, 10, and 10, with $\alpha = 3.21$, 2.5, 1.7, and 1.0. Figures 12 - 14 compare the corresponding density curves with the true curve and a histogram of the raw observations. It seems that EIC and $m = 4$ gives about the right amount of smoothing for a small number of terms in the approximation.

Figure 12: (m = 4, EIC)

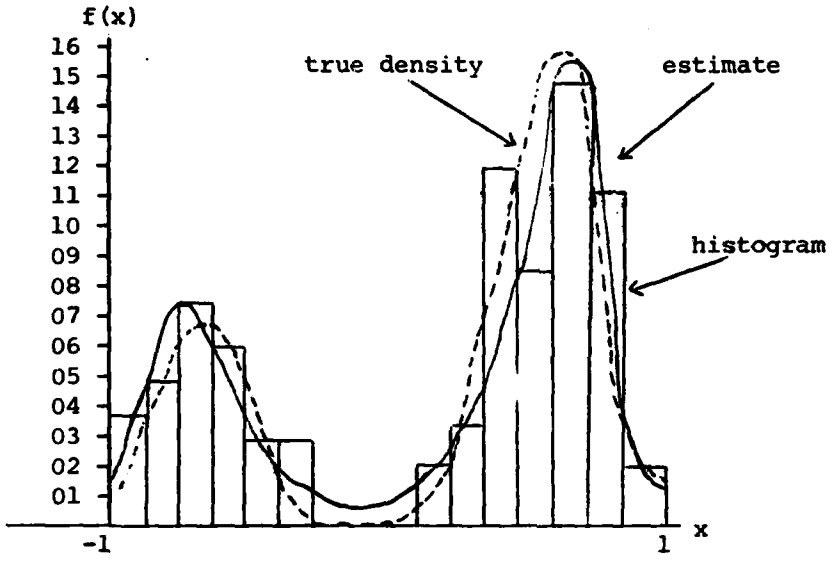


Figure 13: (m = 6, SIC)

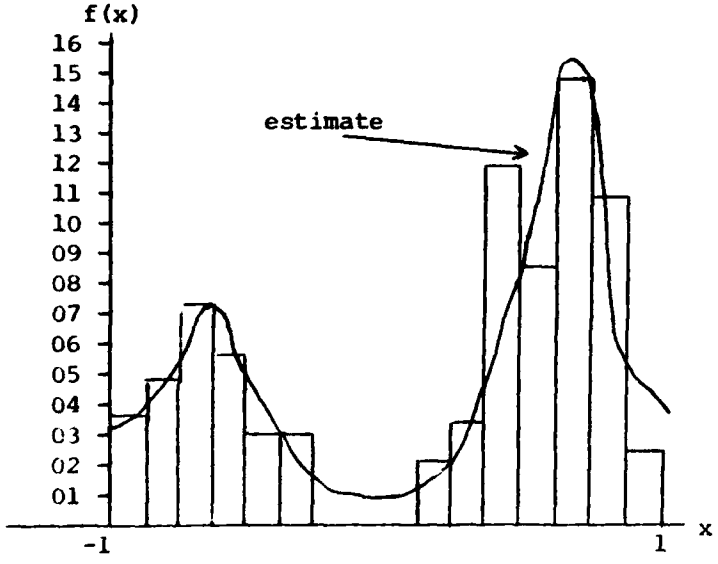
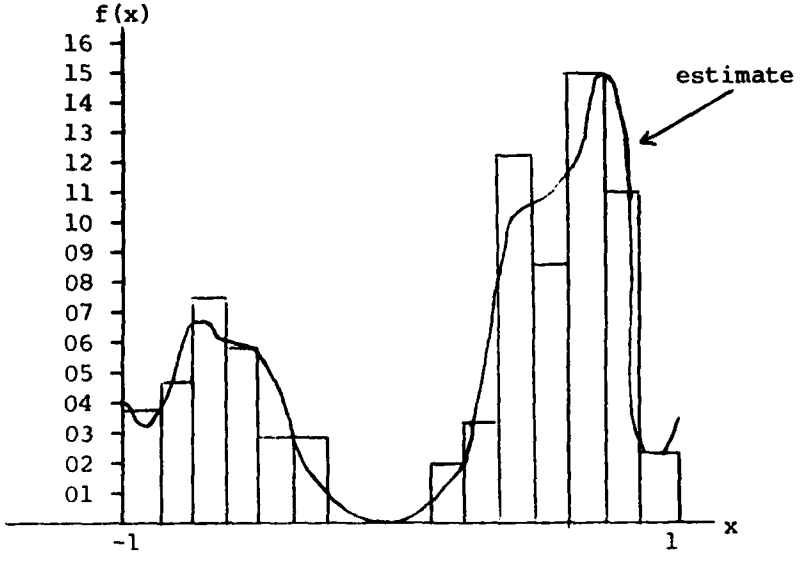


Figure 14: (m = 10, AIC, SIC')



2.3 Basis Selection

It is possible to use our penalized log-likelihoods to compare different choices of basis functions. In the following example we illustrate this approach, as part of a more general simulation study, by comparing Hermite and Laguerre functions.

We completed 500 repeated simulations, and each time simulated a random sample of size 100 from a chi-squared distribution with 14 degrees of freedom. In each case we used the logistic method described in section 2.2, to estimate the density, and the 500 repetitions enabled us to get an accurate estimate (IMSE) together with its standard deviation (STND), and average values for the log-likelihoods together with AIC, SIC, and SIC'. The results are summarized in Tables 4 and 5.

From the values of m which we were able to calculate we see that the preference for Laguerre functions is very clear under IMSE, AIC, SIC, and SIC', giving useful insight which would be difficult to judge intuitively.

ACKNOWLEDGEMENTS

The authors wish to thank Chien F. Wu, Grace Wahba, and Dennis Cox for helpful comments.

Table 4: Simulated Results (Hermite Functions)

m	IMSE	STND	L_m	AIC	SIC	SIC'
2.	.01206	.00066	-3113	-3115	-3119	-3117
4.	.01206	.00066	-2819	-2823	-2831	-2828
6.	.01207	,00066	-2560	-2566	-2578	-2573
8.	.01207	.00066	-2341	-2349	-2366	-2359
10.	.01208	.00066	-2116	-2126	-2147	-2138
12.	.01209	.00065	-1964	-1976	-2002	-1991
14.	.01209	.00065	-1820	-1834	-1863	-1850

Table 5: Simulated Results (Laguerre Functions)

m	IMSE	STND	L_M	AIC	SIC	SIC'
2.	.01127	.00071	-584	-586	-588	-587
4.	.00391	.00042	-301	-305	-310	-307
6.	.00220	.00023	-276	-282	-290	-285
8.	.00212	.00025	-273	-281	-291	-284
10.	.00211	.00026	-270	-280	-293	-284
12.	.00211	.00026	-276	-288	-303	-292
14.	,00211	,00026	-264	-278	-296	-283

REFERENCES

Ahrens, L. H. (1965). "Observations on the Fe-Si-Mg Relationship in Chondrites", Geochiemica et Cosmochiemica Acta, 29, 801-806.

Akaike, H. (1974). "A New Look at the Statistical Model Identification", IEEE Trans. Automat. Contr., AC-19, 716-723.

_____ (1981). "Statistical Inference and Measurement of Entropy", presented at the Conference on Scientific Inference, Data Analysis, and Robustness, November 4-6, 1981.

Basharin, G. P. (1959). "On Statistical Estimate for the Entropy of a Sequence of Independent Random Variables:, in N. Artin (ed.), Theory of Probability and Its Applications, Vol. IV, (Translation of Teoriya Veroyatnostei i ee pvineniya) Society for Industrial and Applied Mathematics, Philadelphia, pp. 333-336.

Box, G. E. P., Kanamasu, H. (1973). "Posterior Probabilities of Candidate Model Discrimination." Technical Report #322, University of Wisconsin.

Chow, G. C. (1981). "A Comparison of the Information and Posterior Probability Criteria for Model Selection", Journal of Econometrics, 16, 21-23.

Crain, B. R. (1974). "Estimation of Distributions Using Orthogonal Expansions", The Annals of Statistics, Vol. 2, #3, 454-463.

Geisser, S., Eddy, W. F. (1979). "A Predictive Approach to Model Selection", J.A.S.A. 74, 153-160.

Good, I. J., Gaskins, R. A. (1980). "Density Estimation and Bump Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data", J.A.S.A., 75, 42-56.

_____ (1971). "Nonparametric Roughness Penalties for Probability Densities", Biometrica, 58, 2, p. 255.

Jeffreys, H. (1961). Theory of Probability, 3rd ed. (Clarendon, Oxford).

Leonard, T. (1973). A Bayesian Method for Histograms. Biometrika, 60, 297-308.

_____ (1978). "Density Estimation, Stochastic Processes, and Prior Information", J.R.S.S., Ser. B, 40, 113-146.

_____ (1981). Comment on the paper by Lejeune and Faulkenberry. To appear in JASA.

Mallows, C. L. (1973). "Some Comments on C_p", Technometrics, 15, 661-675.

Schwarz, G. (1978). "Estimating the Dimension of a Model", Annals of Statistics, 6, 461-464.

Silverman, B. W. (1981). "Using Kernel Density Estimates to Investigate Multimodality", J.R.S.S., Ser. B, 43, 97-99.

Stone, M. (1977). "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike Criterion", J.R.S.S., Ser. B, 39, 44-47.

Wahba, G (1977). "Optimal Smoothing of Density Estimates", Academic Press, pp. 423-457.

TA/TL/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2336	2. GOVT ACCESSION NO. AD-A114 623	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Some Penalized Likelihood Procedures for Smoothing Probability Densities		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Taskin Atilgan and Tom Leonard		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE February 1982 13. NUMBER OF PAGES 25
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Non-parametric, Density estimation, Histogram, Information Criteria, Consistency.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Some methods are considered for the estimation of probability densities. They employ a linear approximation to either the density or the logistic density transform. The coefficients in the approximation are estimated by maximum likelihood and the number of terms is judged via an information criterion. Hence the traditionally difficult problem of judging the degree of smoothness is carried out in a relatively simple manner. Criteria considered include the penalties proposed by Akaike and Schwarz for model complexity, together with an empirical criterion based upon a plot of the log-likelihoods. The practical procedures are related to an asymptotic consistency argument and a number of numerical examples are presented.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)